Processing of Reverberant Speech for Time-Delay Estimation

B. Yegnanarayana, *Senior Member, IEEE*, S. R. Mahadeva Prasanna, *Member, IEEE*, Ramani Duraiswami, *Member, IEEE*, and Dmitry Zotkin, *Member, IEEE*

***Abstract*—In this paper, we present a method of extracting the time-delay between speech signals collected at two microphone locations. Time-delay estimation from microphone outputs is the first step for many sound localization algorithms, and also for enhancement of speech. For time-delay estimation, speech signals are normally processed using short-time spectral information (either magnitude or phase or both). The spectral features are affected by degradations in speech caused by noise and reverberation. Features corresponding to the excitation source of the speech production mechanism are robust to such degradations. We show that these source features can be extracted reliably from the speech signal. The time-delay estimate can be obtained using the features extracted even from short segments (50–100 ms) of speech from a pair of microphones. The proposed method for time-delay estimation is found to perform better than the generalized cross-correlation (GCC) approach. A method for enhancement of speech is also proposed using the knowledge of the time-delay and the information of the excitation source.**

***Index Terms*—Hilbert envelope, source features, speech enhancement, time-delay.**

I. Introduction

LOCALIZATION of the source speaker from the speech signal collected at an array of microphones and enhancement of the received speech signals are challenging tasks [1], [2]. While the same speech information is present at all the microphones in the array, the received signal is different at different microphones. The nature and quality of the received signal depends not only on the distance of travel of the direct sound from the speaker, but also on several other factors, such as other speaker's speech, background noise, and distortion due to multiple reflections or reverberation. The received speech signals from pairs of microphones are processed to estimate the time-delay between each pair. The estimated time-delay information is useful to derive the location of the source speaker. The time-delay information is also useful to enhance the speech, by combining the output signals from several microphones after compensating for the time-delays.

The problems of time-delay estimation, localization of single/multiple speaker sources and enhancement of degraded speech, all have been handled traditionally by exploiting the spectral characteristics of speech signals [3], [4]. The three broad strategies used in these studies are [5] 1) Steered response power of a beam-former, 2) high resolution spectrum estimation, and 3) time difference of arrival estimation. In the steered beam-former the microphone array is steered to various locations to search for a peak in the output power. The delay and sum beam-former time shifts the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. Sophisticated beam-formers apply filtering to the array signals before time alignment and summing. These beam-formers depend on the spectral content of the source signal. *A priori* knowledge of the independent background noise is used to improve the performance [6].

The second category of source locators are based on high resolution spectrum estimation. In this case the spatio-spectral correlation matrix is derived from the signals received at the microphones. This matrix is derived using an ensemble average of the signals over the intervals in which the noise and the sources are assumed to be stationary, and their estimation parameters are assumed to be fixed [7]. These high resolution methods are designed for narrowband stationary signals, and hence are difficult to apply in the case of wideband nonstationary signals like speech.

Methods based on the time differences of arrival (TDOA) estimation are more suitable for speech source localization than the previous two approaches [5]. These involve a two-step process: 1) Obtaining the time-delay estimation of the speech signals corresponding to a pair of spatially separated microphones and 2) using the estimated delays and knowledge of the microphone separation for determining the location of the source. For estimation of the time-delays, the weighted generalized cross-correlation (GCC) method is often used [8]. The method relies on the spectral characteristics of the signal. Since the spectral characteristics of the received signal are modified by the multipath propagation in a room, the GCC function is made more robust by deemphasizing the frequency-dependent weightings [9]. The phase transform is one extreme where the magnitude spectrum is flattened. However then the low signal-to-noise ratio (SNR) portions of the spectrum are given equal emphasis as the high SNR portions. Cepstral prefiltering, which was proposed to reduce the effects of reverberation, is also difficult to apply to speech signals due to the nonstationary

Manuscript received May 3, 2002; revised August 16, 2004. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng. R. Duraiswami and D. Zotkin were supported by NSF Award 0086075; B. Yegnanarayana was supported in part by the same award.

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India (e-mail: yegna@cs.iitm.ernet.in).

S. R. Mahadeva Prasanna is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, North Guwahati 781039, Assam, India (e-mail: prasanna@iitg.ernet.in).

R. Duraiswami and D. Zotkin are with Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: ramani@umiacs.umd.edu; dz@umiacs.umd.edu).

Digital Object Identifier 10.1109/TSA.2005.853005

nature of the signal [10]. Moreover, this approach is not suitable for estimating the time-delays from short (50–100 ms) segments.

Recently methods have been suggested for speech source localization using explicit modeling of speech [11]. Strategies are proposed separately for both the frequency domain and time domain models. In the frequency domain, models with explicit knowledge of the periodicity of the voiced speech are considered to improve the spectrum estimation from the received speech signals [12]. Time domain methods use the knowledge of the large prediction error at the closure of the glottal vibration cycle in the voiced segments. The error due to reverberant components occurs at different random instants for different microphone locations. This information, together with the knowledge of the time-delays, is used to determine a weight function that emphasizes the residual around the instants of glottal closure. This is used to estimate speech from the degraded signal.

It is obvious that most methods for time-delay estimation and source localization rely on the spectral characteristics of the source signal, and the knowledge of the degrading noise and environment. The spectrum of the received signal depends on how the waveform is modified due to distance, noise and reverberation. Therefore the spectra of the signals obtained at two different microphone locations differ significantly. Compensating for the spectral side effects or enhancement of the speech spectral components have met with limited success, as there still will be a lack of coherence in the filtered or spectral compensated signals from different microphones.

In this paper we propose a method that relies on some features of the excitation source of voiced speech for estimating the time-delays [13]–[20]. The method is based on exploiting the characteristics of excitation source especially for voiced speech. The excitation source for voiced speech consists of impulse-like excitation around the instants of glottal closure, called the epochs. The sequence of these impulse-like excitation are robust to degradation in the sense that the relative spacing of these epochs due to direct sound remain unchanged at different microphone locations. On the other hand, the impulse-like excitation due to reflected sound occurs at random locations at the microphones. In this work the impulse-like excitation characteristics are captured by using the Hilbert envelope of the linear prediction (LP) residual of voiced speech. The Hilbert envelopes can be added coherently to reinforce the direct sound and reduce relatively the effects of noise and reverberation. Thus the proposed method is better than the previous efforts because: 1) the vocal tract influence which changes more rapidly, is removed, 2) the Hilbert envelope tends to emphasize the instants of significant excitation, and 3) the instants of significant excitation from the Hilbert envelopes of different microphones can be cohered to get a more reliable output. For coherent addition of the Hilbert envelopes the time-delay between two microphone signals need to be estimated.

This paper is organized as follows. In Section II, we discuss the characteristics of the speech signal that can be exploited for developing the proposed nonspectrum-based approach for time-delay estimation. The implementation of the new method for estimating the time-delays is given in Section III. In Section IV, the performance of the proposed method is compared with the GCC approach. Knowledge of the time-delay estimation can be used to enhance the speech from a degraded signal. An approach for enhancement of speech is presented in Section V. Finally, in Section VI a summary of the work is given along with some issues for further study.

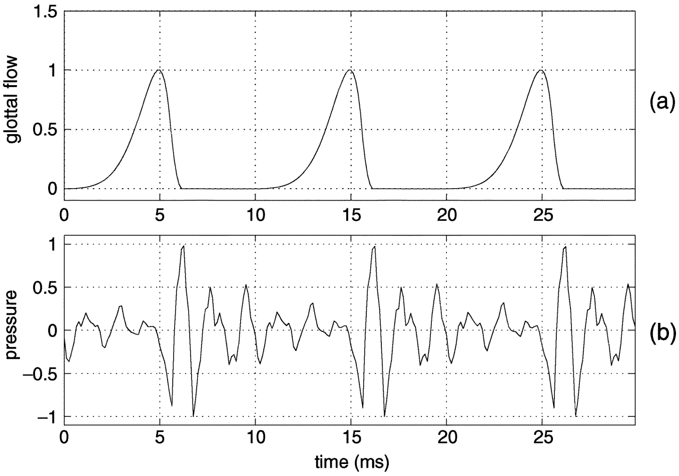

Fig. 1. Nature of excitation of voiced speech: (a) glottal volume velocity and (b) speech waveform.

II. Nature of the Speech Signal

In our proposed method for time-delay estimation the production characteristics of speech are exploited to extract the relevant information from the degraded speech signal received at a microphone. Speech is the result of exciting a time-varying vocal tract system with a time-varying excitation. The common and significant mode of excitation of the vocal tract system is the voiced excitation caused by the vibrating vocal folds at the glottis. The significant excitation of the vocal tract system takes place at the instant of glottal closure within each glottal cycle [18]. The rate of closure is an indication of the strength of excitation. In speech the response of the vocal tract system is superimposed on the sequence of the glottal excitation pulses. The response of the vocal tract system depends on the shape the vocal tract takes to produce a given sound. The resonance characteristics of the vocal tract system appear as damped-sinusoid-like waveforms within each glottal cycle (see, Fig. 1). Since the waveform is affected by the transmission medium, noise and the response of the room, the received speech signal contains information about the vocal tract system corrupted by different types of degradations at different microphones. It is difficult to determine the characteristics of these degradations to compensate for their effects by processing the received signal.

The instants of significant excitation in a voiced segment are unique, and their locations along the time scale do not vary with the transfer characteristics of the medium and the microphones [18]. The identification of the epochs due to direct sound is difficult due to presence of reverberation component in the speech signal. It is important to note that the effect of reverberation is different in different regions of a voiced segment [19]. For example, in the vowel region of a typical syllable-like unit the initial high energy pitch periods are less affected by reverberation compared to the later pitch periods.

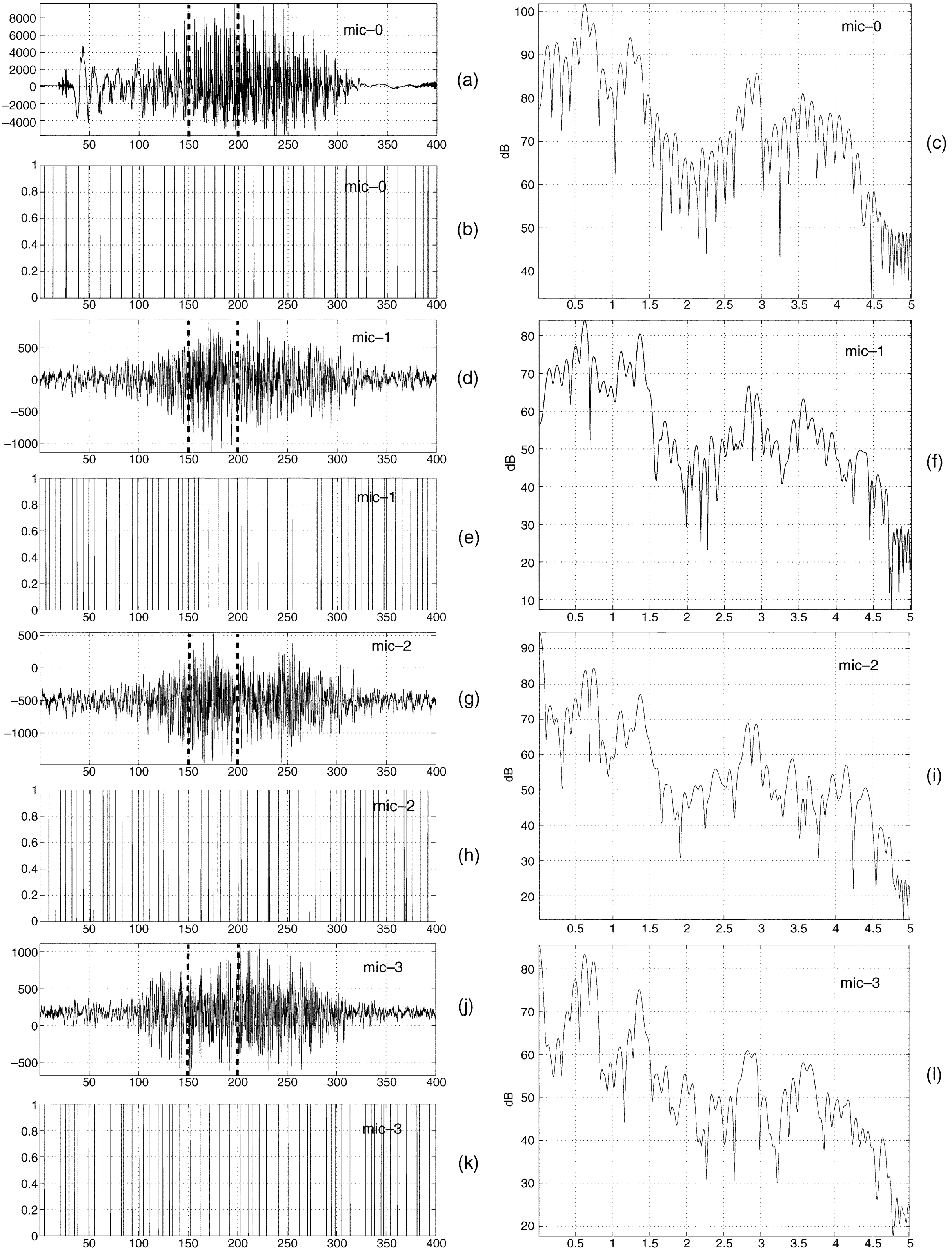

Fig. 2. Nature of speech signals at four different microphone locations (*mic-0*, *mic-1*, *mic-2*, and *mic-3*). (a), (d), (g), and (j) Waveforms of the speech segments at the four microphone locations. (b), (e), (h), and (k) are the extracted instants of significant excitation corresponding to the four segments. (c), (f), (i), and (l) are the short-time spectra for the portions marked in the segments.

Fig. 2(a), (d), (g) and (j), respectively, shows the signals received by a close speaking microphone (*mic-0*) and 3 other microphones (say, *mic-1*, *mic-2* and *mic-3*) placed in a normal room (3 m × 4 m × 3 m) with an average reverberation time of about 200 ms. The waveforms are clearly different from each other, and from the clean speech waveform obtained with a

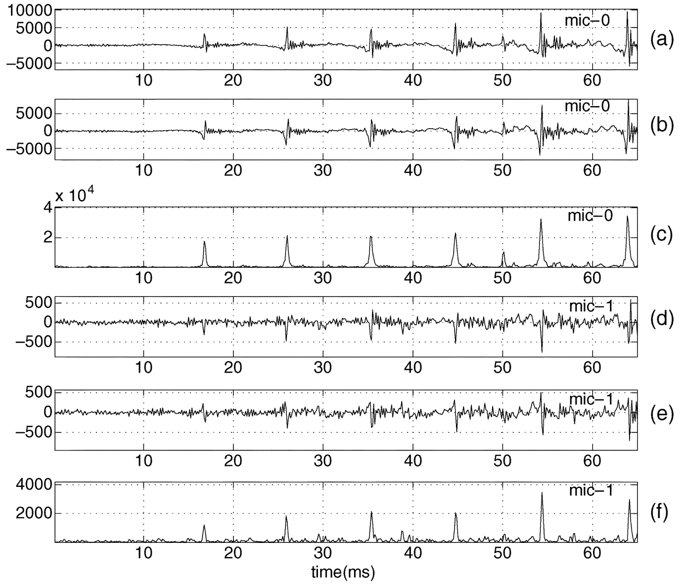

Fig. 3. Illustration of the characteristics of the Hilbert envelope of a voiced speech segment. (a), (b) and (c) are respectively, the LP residual, its Hilbert transform, and the Hilbert envelope for the signal at *mic-0*. (d), (e), and (f) are respectively, the LP residual, its Hilbert transform and the Hilbert envelope for the signal at *mic-1*.

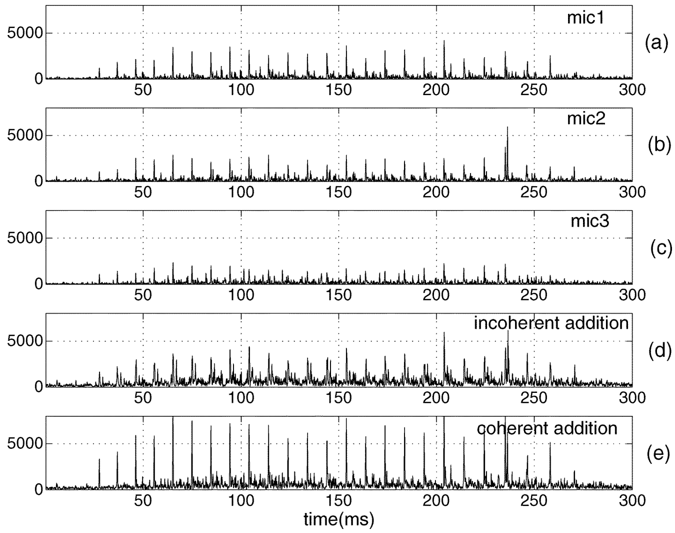

Fig. 4. Effect of coherent and incoherent additions of the Hilbert envelopes. (a) Hilbert envelope for the signal at *mic-1*. (b) Hilbert envelope for the signal at *mic-2*. (c) Hilbert envelope for the signal at *mic-3*. (d) Result of *incoherent* addition of the Hilbert envelopes. (e) Result of *coherent* addition of the Hilbert envelopes.

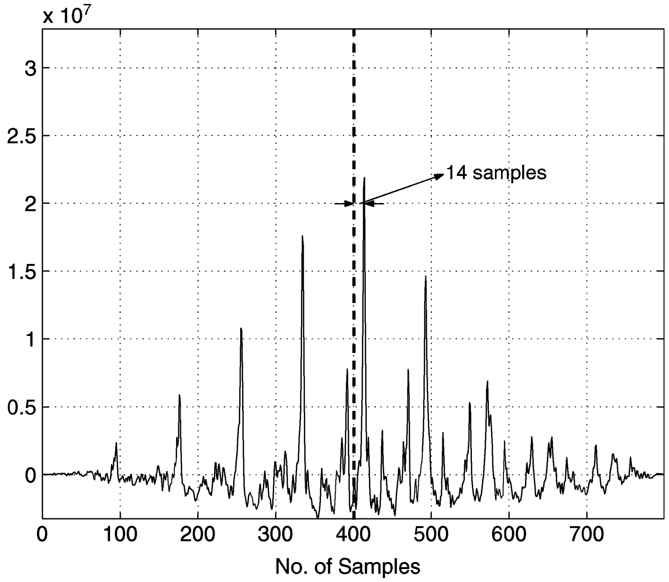

Fig. 5. Cross-correlation function of the Hilbert envelopes of frames of size 50 ms (400 samples), corresponding to signals at *mic-1* and *mic-2*. The time-delay between *mic-1* and *mic-2* signals is 14 samples.

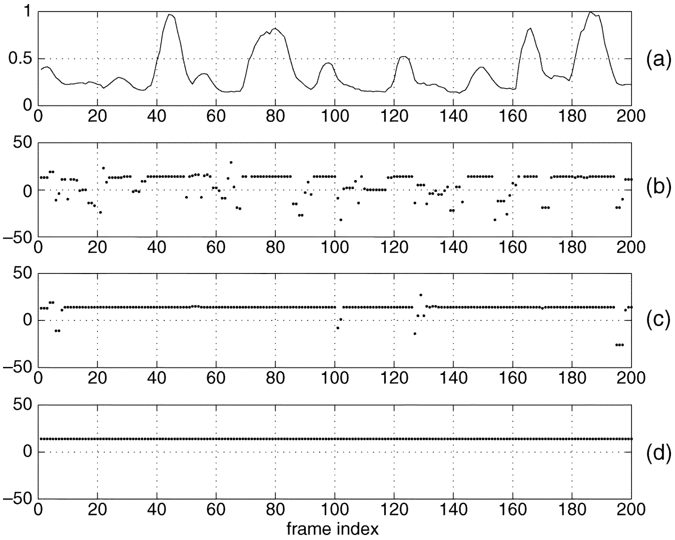

Fig. 6. Characteristics of the estimated time-delay for different sizes of analysis frame. (a) Normalized Hilbert envelope energy. Time-delays from analysis frames of size (b) 50 ms, (c) 200 ms, and (d) 500 ms, each with a shift of 10 ms.

close speaking microphone. The figure also shows the short-time (50 ms frame shown by dashed lines) spectra for each of the segments to show the differences in the short-time spectral envelopes. Fig. 2(b), (e), (h) and (k) shows the epoch locations for all the four cases of microphone signals using the method proposed in [15]. The epochs are computed from the LP residual of the received signal. The LP residual is obtained using a 10th-order LP analysis [21]. Throughout this study the speech signals sampled at 8 kHz are used. It is obvious from the clean speech case [Fig. 2(b)] that if the epoch locations can be derived from the received signals, the problem of time-delay estimation is not only trivial, but also the resulting estimation will be accurate. But the spurious epochs due to noise and reverberation makes it difficult to use the epoch locations directly for time-delay estimation.

A better method to estimate the time-delay is to exploit the property that the strength of excitation in voiced speech is large around the glottal closure instant. The impulse-like excitation results in large error in the LP residual around the glottal closure instant. But it is difficult to determine the location of the glottal closure instant due to fluctuations of the amplitudes of the residual samples depending on the phase of the signal. Ideally it is desirable to derive an impulse-like signal around the glottal closure instant. A close approximation to this is possible by using the Hilbert envelope of the LP residual, instead of the energy in short intervals of time. Even though the real and

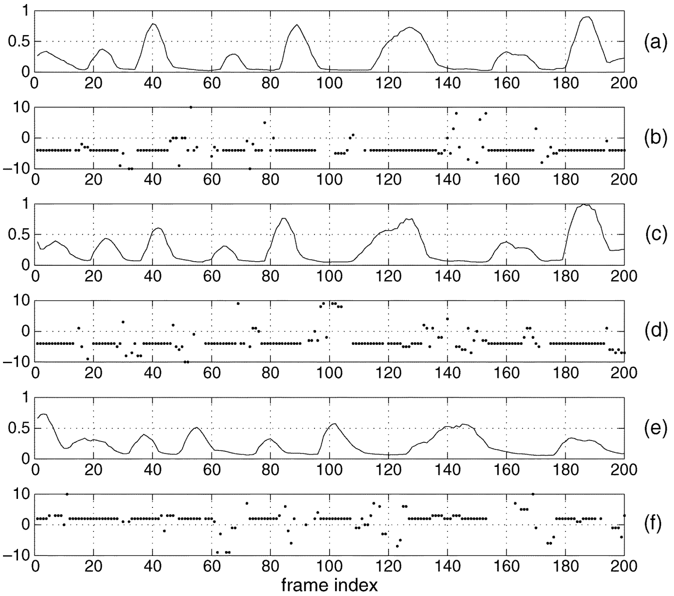

Fig. 7. Characteristics of the estimated time-delay for different levels of reverberation obtained by collecting speech data at distances (b) 2 ft, (d) 4 ft, and (f) 6 ft from the speaker. The time-delays are estimated using an analysis frame of 50 ms with a shift of 10 ms. The corresponding normalized Hilbert envelope energy plots are shown for each case in (a), (c), and (e), respectively.

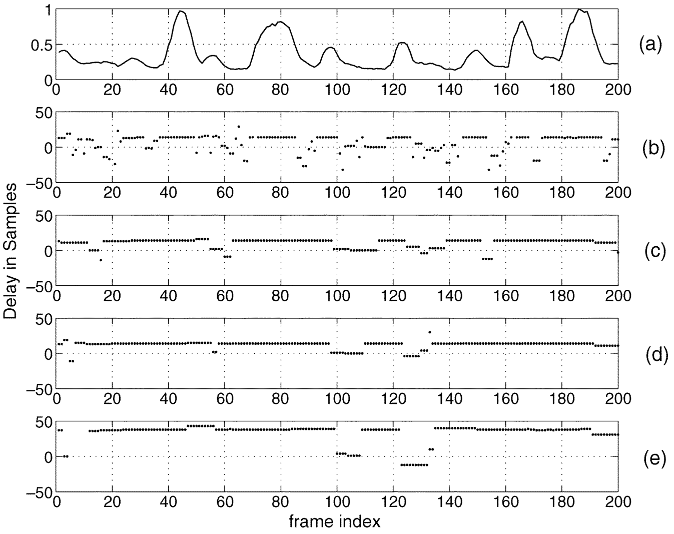

Fig. 8. Characteristics of the estimated time-delay for speech degraded by different types of noises, namely, (a) air-conditioning noise, (b) fan noise and (c) both air-conditioning and fan noise collected at a distance of 6 ft from the speaker. The microphones were placed close to the noise source. The time-delays are estimated using an analysis frame of 50 ms with a shift of 10 ms. The normalized Hilbert envelope energy plots are shown for each case in (a), (c), and (e), respectively.

imaginary parts of an analytic signal (related through the Hilbert transform) have positive and negative samples, the Hilbert envelope of the signal is a positive function, giving the envelope of the signal [22]. For example, the Hilbert envelope of a unit sample sequence or its derivative has a peak at the same instant. Thus the properties of Hilbert envelope can be exploited to derive the impulse-like characteristics of the excitation. The Hilbert envelope $h(n)$ of signal $e(n)$ is defined as follows [14], [22], [23]:

$$h(n) = \sqrt{e^2(n) + e_H^2(n)} \tag{1}$$

where $e_H(n)$ is the Hilbert transform of $e(n)$, and is given by [14]

$$e_H(n) = \begin{Bmatrix} IDFT\,[jE(\omega)], & -\pi < \omega \leq 0 \\ IDFT\,[-jE(\omega)], & 0 < \omega \leq \pi \end{Bmatrix} \tag{2}$$

where DFT and IDFT are the discrete and inverse discrete fourier transforms, respectively, and $E(\omega)$ is the DFT of $e(n)$.

Fig. 3 shows the LP residual, its Hilbert transform and the Hilbert envelope for a segment of the speech signal at the close speaking microphone (*mic-0*) and also for a segment of the degraded speech signal at *mic-1*. The figure clearly illustrates the important property of the Hilbert envelope of a voiced speech segment, namely, the peak of the envelope occurs around the instant of glottal closure within each pitch period. Note that due to phase relations among the signal samples, the LP residual or its Hilbert transform need not have a peak at the instant of glottal closure. This can be seen at each instant of the glottal closure in the plots in Fig. 3(a)–(c) for the clean speech signal from *mic-0*. Even for the degraded speech signal at *mic-1*, the Hilbert envelope shows the largest peak around the instant of glottal closure within each pitch period. This important property of the Hilbert envelope forms the basis for the method proposed in this paper for estimating the time-delay.

While the amplitude of the Hilbert envelope is high at the instant of significant excitation, the amplitudes of the Hilbert envelope will also be high at the epochs of the reflected sound in the reverberant speech. But these epochs will be located at random instants. In the next section we will show how the Hilbert envelopes of the LP residual signals from different microphones can be used to estimate the time-delay for each pair of microphones.

III. Estimation of Time-Delay From Hilbert Envelope of the LP Residual

The Hilbert envelope of the LP residual signal shows large amplitudes around the instants where the residual error is large. The error is large around the instants of glottal closure in the direct signal as well as in the reverberant components of the signal. The instants corresponding to the direct signal will be *coherent* at different microphone positions. On the other hand, the instants corresponding to the reverberation components will be at random locations along the time scale. This can be seen from Fig. 4, where the Hilbert envelopes for signals from the three microphone positions are time aligned and displayed. The effect of coherence of the direct components can be seen when we add the delay-compensated Hilbert envelope signals from the three microphones. It is important to note that the coherent addition in Fig. 4(e) produces significant peaks at the epochs, whereas the incoherent addition in Fig. 4(d) produces several peaks at random locations.

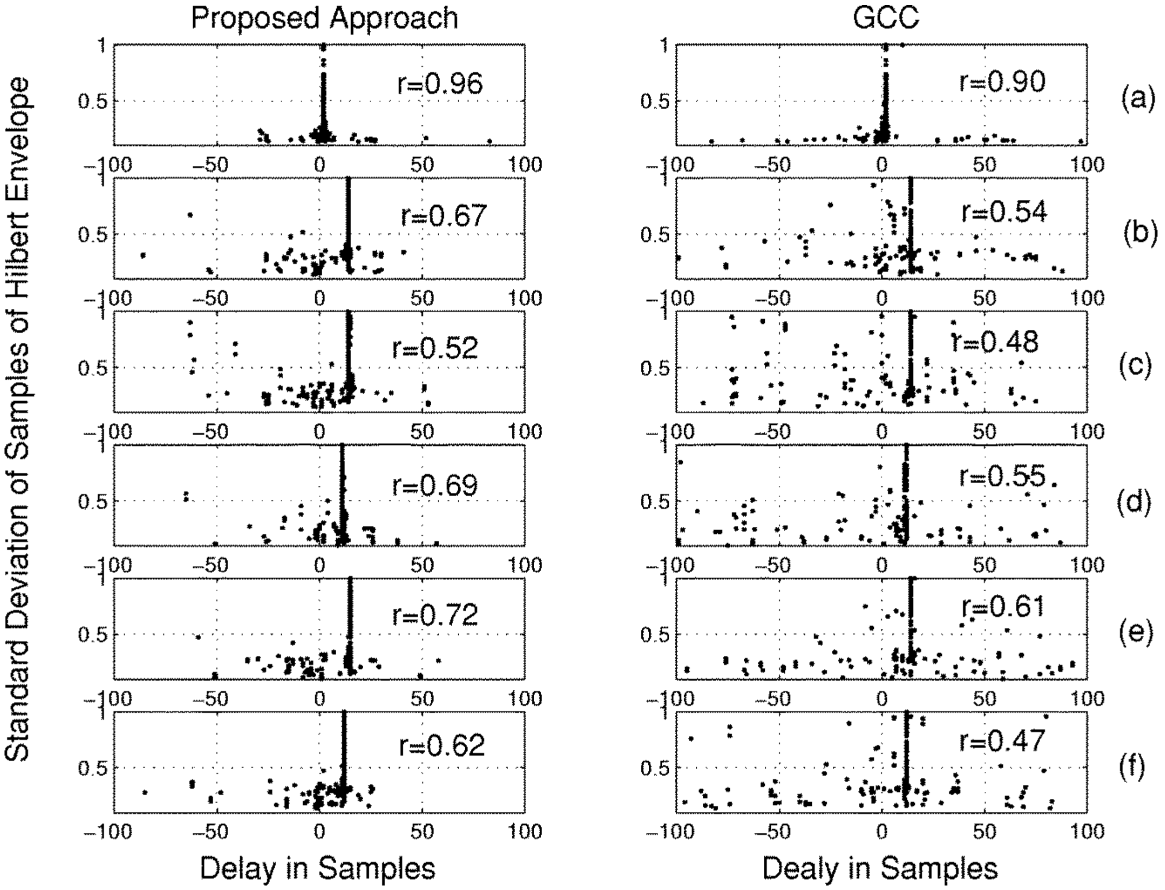

Fig. 9. Quantitative comparison of the proposed time-delay estimation method with GCC. Standard deviation of the samples of the Hilbert envelope vs. estimated time-delay (in samples) are shown for different cases. (a) *mic-1* and *mic-3* signals. (b) *mic-1* and *mic-2* signals. (c) *mic-2* and *mic-11* signals. (d) *mic-1* and *mic-14* signals. (e) *mic-1* and *mic-15* signals. (f) *mic-2* and *mic-9* signals.

For the coherent addition one needs the values of the time-delays. We propose a cross-correlation method to determine the time-delays. Consider a frame of 50 ms from one of the microphones, say *mic-1*, and compute the cross-correlation of the Hilbert envelope of the LP residual of this frame and the corresponding frame of 50 ms from the second microphone, say *mic-2*. The location of the peak in the cross-correlation corresponds to the delay. The time-delay to be estimated is assumed to be much less (<10%) than the size of the frame (50 ms in this case) being considered. Fig. 5 shows the cross-correlation function of the Hilbert envelopes of segments of the two microphone signals. The delay is indicated in number of samples from the center sample number, which is 400 in this case.

The time-delay for each frame of 50 ms is computed with a shift of 10 ms between successive frames, and the result is plotted in Fig. 6(b). Note that the estimation results in random delays, mostly for nonvoiced segments. This is indicated by segments corresponding to the low energy regions of the Hilbert envelope, as shown in Fig. 6(a). The energy of the Hilbert envelope is obtained for each frame of 50 ms by computing the mean squared values of the amplitudes of the envelope within the frame. The normalized energy plot in Fig. 6(a) is obtained by computing the energy for each frame shifted by 10 ms, and normalizing the energy values by dividing them with the maximum value over the segment. The regions of the low normalized values, say values below 0.25 in the Fig. 6(a), correspond mostly to silence or noise or unvoiced or low voicing regions. The low voicing regions are those regions where the strength of excitation around the glottal closure is not high. This is also indicated by the lower values of the Hilbert envelope relative to the values in the high voicing regions.

The estimation of the time-delay gets better when we consider longer frame sizes as shown in Fig. 6(b) and (c), for frame sizes of 200 ms and 500 ms, respectively. The improvement in the delay estimate is indicated by fewer spurious or random delays compared to the case of 50 ms frame. But using longer segments for delay estimation may make it difficult to keep track of a moving source/speaker.

Figs. 7 and 8 illustrate the performance of the time-delay estimation using the Hilbert envelope of the LP residual for two different types of degradation. The time-delay estimation for different levels of reverberation are obtained by placing the microphones at different distances from the speaker. Note that for longer distances the SNR decreases due to the constant background noise. In Fig. 8 the estimated time-delays are

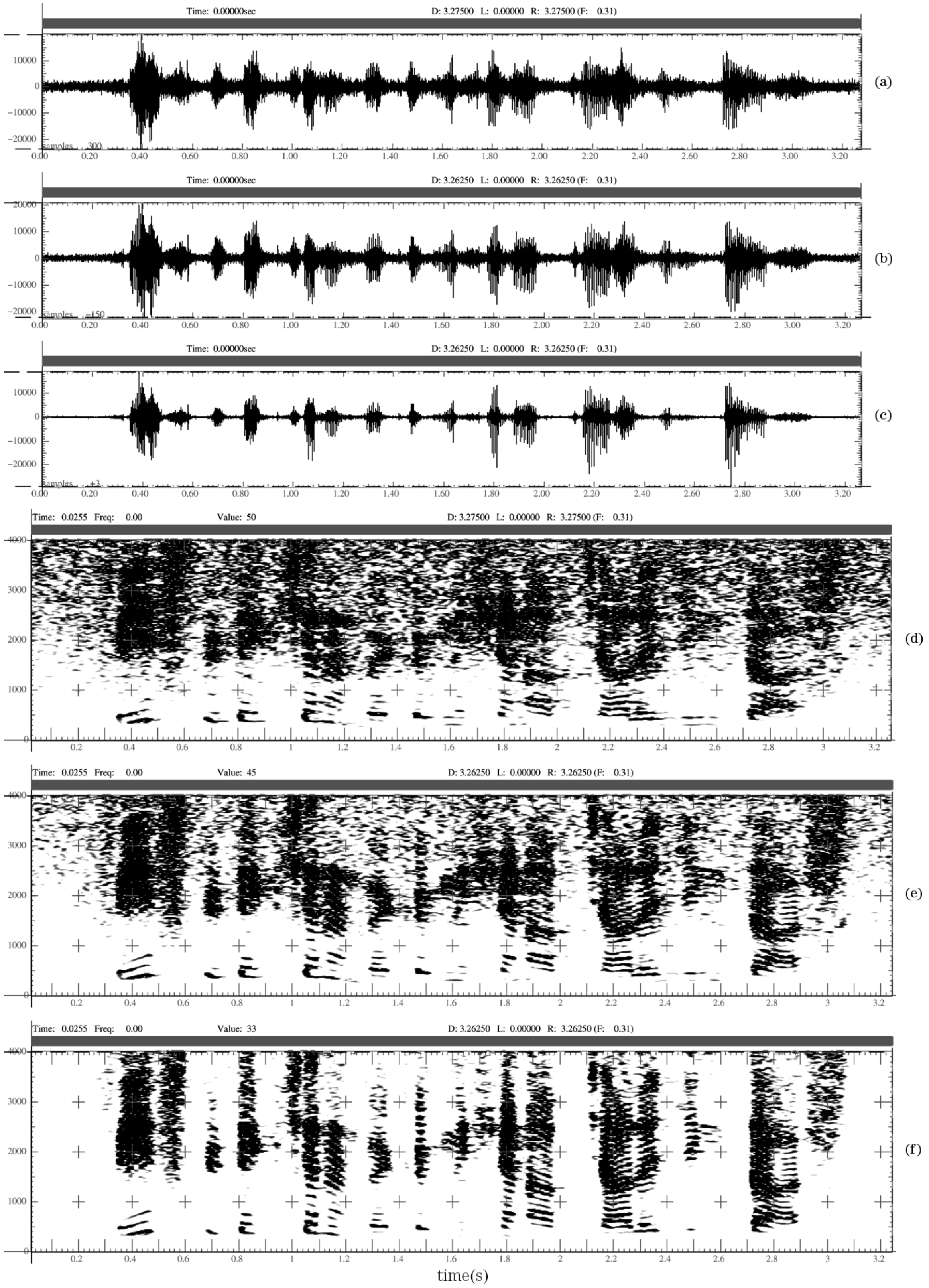

Fig. 10. Results of enhancement of reverberant speech. (a) Degraded speech. (b) Enhanced speech by waveform addition. (c) Enhanced speech by the proposed method. (d) Spectrogram of the degraded speech. (e) Spectrogram of the enhanced speech obtained by waveform addition. (f) Spectrogram of the enhanced speech obtained by the proposed method.

plotted when the microphone is placed close to a room air conditioner and a fan. A constant value of the time-delay is obtained for successive frames in the regions of high energy values in all the plots in Figs. 7 and 8.

IV. Comparison With the GCC Method

In this section we compare the time-delays estimated by the proposed nonspectral approach with the standard GCC approach. The GCC $R_{x_1x_2}(\tau)$ is computed as the inverse Fourier transform of the cross-spectrum $X_1(\omega)X_2^*(\omega)$ of the received signals, scaled by a weighting function $W(\omega)$ [9]. That is

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} W(\omega)X_1(\omega)X_2^*(\omega)e^{j\omega\tau}d\omega \tag{3}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the microphone signals $x_1(t)$ and $x_2(t)$. The weight function is chosen as $W(\omega) = |X_1(\omega)X_2^*(\omega)|^{-1}$. This corresponds to the use of phase transform for cross-correlation.

Since accurate estimation of the time-delays with smaller frame size helps in tracking a moving source, we consider a frame of 50 ms with a shift of 10 ms to compare the performance of the time-delay estimation by the proposed method and by the GCC method. Note that in the Hilbert envelope plot in Fig. 3, for the voiced segments there are large values around the instants of glottal closure, followed by small values within each pitch period. Therefore the variance of the sample values of the Hilbert envelope in the voiced region will be large, compared to the nonvoiced segments where the sample values are relatively more uniformly distributed, thus contributing to low variance. The variance of the sample values of the Hilbert envelopes is computed for each frame of 50 ms shifted by 10 ms. Plots of the standard deviation of the samples of the Hilbert envelope against estimated time-delays for every frame of 50 ms with a shift of 10 ms are shown in Fig. 9. The figure shows the plots for different pairs of microphones to illustrate the effects of degradations as the signals at different microphones are not of the same quality. One important point to be noted from these plots is that, even in the low voiced regions (low standard deviation), the proposed method estimates the delays accurately, whereas the GCC shows significant variations in the estimated delays. Ideally all the points should lie along a vertical line at the delay value. So the spread of points from the vertical line indicates degradation in the performance of the method.

An objective measure for comparison of the performance could be the ratio (r) of the number of points around the time-delay within ± 1 sample deviation to the total number of points above a certain threshold of the value of the standard deviation of the samples of the Hilbert envelope. Since lower values of the standard deviation correspond mostly to nonvoiced regions, we can ignore the values below 0.25 for computing this ratio. The values of r for the different cases are shown in Fig. 9. The larger the value of r, the better is the method for estimating the time-delay. From these illustrations we can infer that the proposed method is superior to the GCC method.

V. Enhancement of Speech Using Time-Delay Information

We propose a simple but effective method of enhancing the speech in the received degraded signal. Speech signals from multiple channels are coherently added using the estimated delays. This addition reduces the effect of background noise, but the effects of reverberation will still remain. This can be reduced as follows: The coherent addition of the Hilbert envelopes from different microphone signals yield an envelope function in the time domain as shown in Fig. 4(e). A smoothed envelope function is derived from the coherently-added Hilbert envelope using a running mean of 24 samples corresponding to 3 ms, i.e., less than a pitch period. The resulting envelope function is normalized with respect its maximum value. The square root of the normalized envelope function is used as a weight function for the LP residual derived from the coherently-added signal. The objective of this weighting is to enhance the perceptually significant component of the excitation signal, and reduce the excitation component due to noise and reverberation.

The weighted LP residual is used to excite the all-pole filter derived from the coherently-added speech signal. The degraded speech from one of the microphones, the enhanced speech signal by waveform addition and the enhanced speech signal by the proposed method are shown in Fig. 10. The corresponding spectrograms are also shown in the figure. The enhanced speech sounds significantly better compared to the degraded one. Enhancement of speech can also be obtained using the proposed algorithm for speech degraded by other types of noise, such as reverberation and background noise. Thus the proposed method of time-delay estimation is useful for enhancement of speech degraded by different noise sources. The degraded speech signals and the corresponding enhanced speech signals by the proposed method are available for listening at http://speech.cs.iitm.ernet.in/Main/result/timedelay.html.

VI. Conclusion

In this paper, we have proposed a method for estimating the time-delays from speech signals collected with spatially distributed microphones. The method uses the knowledge of the excitation source, unlike the commonly used spectrum-based methods. The proposed method uses the Hilbert envelope of the LP residual of the speech signal. Since time-delays can be estimated accurately even from short segments of speech, it is also possible to develop algorithms to track a moving speaker.

Knowledge of the time-delays between pairs of microphones can be used to enhance the speech collected at different microphones. This is accomplished by coherently adding the speech signals after compensating for the time-delays. The resulting speech will reduce the background noise. But it still has significant reverberation component. This is reduced by using a weight function for the LP residual of the coherently added signal. The weight function is derived from the coherently added Hilbert envelope of the LP residuals of the speech signals from different microphones.

Several refinements are possible for improving the performance of the proposed enhancement algorithm. A new set of linear prediction coefficients can be derived for each frame using the knowledge of the region around the instants of glottal closure, which is obtained from the coherent sum of the Hilbert envelopes as shown in Fig. 4(e). The new time-varying all-pole filter can be re-excited using the weighted LP residual to obtain further enhancement of speech.

REFERENCES

[1] H. Silverman, W. Patterson, J. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Munich, Germany, Apr. 1997, pp. 251–254.

[2] M. Brandstein, J. Adcock, and H. Silverman, "Microphone array localization error estimation with application to sensor placement," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3807–3816, 1996.

[3] D. Johnson and D. Dudgeon, *Array Signal Processing—Concepts and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[4] G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, pp. 922–926, 1977.

[5] J. H. DiBiase, H. Silverman, and M. Brandstein, *Robust Localization in Reverberant Rooms in Theory and Applications Acoustic Signal Processing for Telecommunications*. Boston, MA: Kluwer, 2000, pp. 131–154.

[6] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 608–614, Sep. 1973.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, pp. 284–290, Mar. 1986.

[8] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 45–50, Jan. 1997.

[9] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, pp. 320–327, Aug. 1976.

[10] A. Stephene and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, Detroit, MI, May 1995, pp. 3055–3058.

[11] M. Brandstein and S. Griebel, *Explicit Speech Modeling for Microphone Array Applications in Theory and Applications Acoustic Signal Processing for Telecommunications*. Boston, MA: Kluwer, 2000.

[12] M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2914–2919, 1999.

[13] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1975.

[14] ——, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[15] B. Yegnanarayana and R. L. H. M. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Detroit, MI, May 1995, pp. 776–779.

[16] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 325–333, Aug. 1995.

[17] B. Yegnanarayana and N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 313–327, Jul. 1998.

[18] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[19] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[20] B. Yegnanarayana, S. R. M. Prasanna, and S. V. Gangashetty, "Autoassociative neural network models for speaker recognition," in *Proc. Int. Workshop Embedded Systems*, Hyderabad, India, 2001.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[22] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[23] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, Orlando, FL, May 2002, pp. 541–544.

B. Yegnanarayana (M'78–SM'84) was born in India in 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1964, 1966, and 1974, respectively

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1978 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai. He was the Chairman of the department from 1985 to 1989. His current research interests are in signal processing, speech, vision, neural networks, and man–machine interfaces. He has published papers in reviewed journals in these areas.

Dr. Yegnanarayana is a Fellow of the Indian National Science Academy, Indian National Academy of Engineering, and Indian Academy of Sciences. He is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

S. R. Mahadeva Prasanna (M'05) was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, India, in 1994, the M.Tech. degree in industrial electronics from National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, in 2004.

He worked as Project Officer in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, from December 2003 to July 2004. Since August 2004, he has been an Assistant Professor in the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, North Guwahati, Assam, India. His research interests are in speech signal processing and neural networks.

Ramani Duraiswami (M'99) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology, Bombay, in 1985 and the Ph.D. degree in mechanical engineering and applied mathematics from The Johns Hopkins University, Baltimore, MD, in 1991.

He is a member of the faculty in the Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park. He is the Director of the Perceptual Interfaces and Reality Laboratory there. His research interests are broad and currently include spatial audio, virtual environments, microphone arrays, computer vision, statistical machine learning, fast multipole methods, and integral equations.

Dmitry Zotkin (M'03) was born in Moscow, Russia, in 1973. He received the B.S. and M.S. degrees in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1996, and the M.S. and Ph.D. degrees in computer science from University of Maryland, College Park, in 1999 and 2002, respectively.

He is currently an Assistant Research Scientist at the Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland. His current research interests are in multichannel signal processing for tracking and multimedia. He is also working in the area of spatial audio, including virtual auditory scene synthesis, customizable virtual auditory displays, perceptual processing interfaces, and associated problems.